Novelty Assessment Report

Paper: The Attacker Moves Second: Stronger Adaptive Attacks Bypass Defenses Against LLM Jailbreaks and Prompt Injections

PDF URL: https://openreview.net/pdf?id=7B9mTg7z25

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2026-01-07

Abstract

How should we evaluate the robustness of language model defenses? Current defenses against jailbreaks and prompt injections (which aim to prevent an attacker from eliciting harmful knowledge or remotely triggering malicious actions, respectively) are typically evaluated either against a static set of harmful attack strings, or against computationally weak optimization methods that were not designed with the defense in mind. We argue that this evaluation process is flawed.

Instead, we should evaluate defenses against adaptive attackers who explicitly modify their attack strategy to counter a defense's design while spending considerable resources to optimize their objective. By systematically tuning and scaling general optimization techniques—gradient descent, reinforcement learning, random search, and human-guided exploration—we bypass 12 recent defenses (based on a diverse set of techniques) with attack success rate above 90% for most; importantly, the majority of defenses originally reported near-zero attack success rates. We believe that future defense work must consider stronger attacks, such as the ones we describe, in order to make reliable and convincing claims of robustness.

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Evaluating Robustness of Language Model Defenses Against Adaptive Attacks**

A total of **50 papers** were analyzed and organized into a taxonomy with **24 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Jailbreak Attack Methods and Characterization**
- **Jailbreak Defense Mechanisms**
- **Adaptive Attack Design and Defense Evaluation**
- **Backdoor Attacks and Defenses in Language Models**
- **Model Extraction and Intellectual Property Protection**
- **Domain-Specific Security and Robustness**
- **General Adversarial Robustness and Defense Techniques**
- **Federated and Distributed Learning Security**

Complete Taxonomy Tree

- Evaluating Robustness of Language Model Defenses Against Adaptive Attacks Survey Taxonomy
- Jailbreak Attack Methods and Characterization
  - Jailbreak Attack Benchmarks and Transferability Analysis (2 papers)
  - [1] Jailbreakv: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks (Weidi Luo, 2024) View paper
  - [5] Jailbreakbench: An open robustness benchmark for jailbreaking large language models (Maksym Andriushchenko, 2024) View paper
  - Multimodal and Cross-Lingual Jailbreak Attacks (2 papers)
  - [26] The Tower of Babel Revisited: Multilingual Jailbreak Prompts on Closed-Source Large Language Models (Huang Linghan, 2025) View paper
  - [29] Visual Adversarial Examples Jailbreak Large Language Models (Qi, 2023) View paper
  - Comprehensive Jailbreak Attack Surveys (3 papers)
  - [17] Jailbreak Attacks and Defenses Against Large Language Models: A Survey (Sibo Yi, 2024) View paper
  - [18] A Comprehensive Study of Jailbreak Attack versus Defense for Large Language Models (Deng Gelei, 2024) View paper
  - [25] From LLMs to MLLMs to Agents: A Survey of Emerging Paradigms in Jailbreak Attacks and Defenses within LLM Ecosystem (Liu Peipei, 2025) View paper
- Jailbreak Defense Mechanisms
  - Input Perturbation and Smoothing-Based Defenses (2 papers)
  - [2] SmoothLLM: Defending Large Language Models Against Jailbreaking Attacks (Robey, 2023) View paper
  - [9] Enhancing robustness of LLM-driven multi-agent systems through randomized smoothing (Hu Jinwei, 2025) View paper
  - Prompt-Based and Interpretable Defense Strategies (3 papers)
  - [7] Defensive Prompt Patch: A Robust and Generalizable Defense of Large Language Models against Jailbreak Attacks (Chen, 2025) View paper
  - [8] {SelfDefend}:{LLMs} can defend themselves against jailbreaking in a practical manner (Wang, 2025) View paper
  - [11] Defensive Prompt Patch: A Robust and Interpretable Defense of LLMs against Jailbreak Attacks (Xiong Chen, 2024) View paper
  - Adversarial Training and Optimization-Based Defenses (3 papers)
  - [6] Robust Prompt Optimization for Defending Language Models Against Jailbreaking Attacks (Bo Li, 2024) View paper

  - [10] Latent-space adversarial training with post-aware calibration for defending large language models against jailbreak attacks (Xin Yi, 2025) View paper
  - [40] Adversarial tuning: Defending against jailbreak attacks for llms (Liu Fan, 2024) View paper
  - Layer-Specific and Mechanistic Defenses (3 papers)
  - [13] Defending large language models against jailbreak attacks via layer-specific editing (Li Zhe, 2024) View paper
  - [19] Beyond surface-level patterns: An essence-driven defense framework against jailbreak attacks in llms (Xiang Shi-yu, 2025) View paper
  - [32] Shieldlearner: A new paradigm for jailbreak attack defense in llms (Ni, 2025) View paper
  - Rapid Adaptation and Few-Shot Defense Learning (1 papers)
  - [43] Rapid Response: Mitigating LLM Jailbreaks with a Few Examples (Michael, 2024) View paper
- Adaptive Attack Design and Defense Evaluation
  - Adaptive Attack Frameworks Against LLM Defenses ★ (2 papers)
  - [0] The Attacker Moves Second: Stronger Adaptive Attacks Bypass Defenses Against LLM Jailbreaks and Prompt Injections (Anon et al., 2026) View paper
  - [44] A Critical Evaluation of Defenses against Prompt Injection Attacks (Jia Yuqi, 2025) View paper
  - Robustness Evaluation Against Character-Level and Textual Perturbations (2 papers)
  - [4] Robustness of Large Language Models Against Adversarial Attacks (Yiyi Tao, 2024) View paper
  - [41] Differential Robustness in Transformer Language Models: Empirical Evaluation Under Adversarial Text Attacks (Ajao, 2025) View paper
  - Baseline Defense Evaluation and Cross-Domain Analysis (2 papers)
  - [15] Baseline Defenses for Adversarial Attacks Against Aligned Language Models (Jain, 2023) View paper
  - [33] Do Methods to Jailbreak and Defend LLMs Generalize Across Languages? (Atil, 2025) View paper
- Backdoor Attacks and Defenses in Language Models
  - Backdoor Defense for Pre-Trained and Few-Shot Language Models (2 papers)
  - [3] Defending pre-trained language models as few-shot learners against backdoor attacks (Xi, 2023) View paper
  - [39] Certifying Language Model Robustness with Fuzzed Randomized Smoothing: An Efficient Defense Against Backdoor Attacks (He, 2025) View paper
  - Clean-Label and Stealthy Backdoor Attack Techniques (2 papers)
  - [23] Exploring Clean Label Backdoor Attacks and Defense in Language Models (SHUAI ZHAO, 2024) View paper
  - [24] Causality based front-door defense against backdoor attack on language models (Y Liu, 2024) View paper
  - Topological and Adaptive Backdoor Defense Strategies (3 papers)
  - [21] TED-LaST: Towards Robust Backdoor Defense Against Adaptive Attacks (Mo Xiaoxing, 2025) View paper
  - [49] Towards Backdoor Attacks and Defense in Robust Machine Learning Models (Soremekun, 2023) View paper
  - [50] TrojanZoo: Towards Unified, Holistic, and Practical Evaluation of Neural Backdoors (Pang Ren, 2022) View paper
- Model Extraction and Intellectual Property Protection
  - Query-Based Model Extraction Attacks and Defenses (3 papers)
  - [12] Prediction Poisoning: Towards Defenses Against DNN Model Stealing Attacks (Orekondy, 2023) View paper
  - [16] Query Provenance Analysis: Efficient and Robust Defense Against Query-Based Black-Box Attacks (Shaofei Li, 2025) View paper
  - [31] A Comprehensive Defense Framework Against Model Extraction Attacks (Wenbo Jiang, 2023) View paper
  - Watermarking and IP Protection for LLMs (1 papers)
  - [22] ModelShield: Adaptive and Robust Watermark Against Model Extraction Attack (Kaiyi Pang, 2025) View paper
- Domain-Specific Security and Robustness
  - Security Hardening for Code Generation Models (2 papers)
  - [20] Large Language Models for Code: Security Hardening and Adversarial Testing (He, 2023) View paper
  - [28] Attacks and Defenses for Large Language Models on Coding Tasks (Chi Zhang, 2024) View paper
  - Multimodal and Speech-Based Adversarial Robustness (2 papers)
  - [34] SpeechGuard: Exploring the adversarial robustness of multimodal large language models (Bhatia, 2024) View paper
  - [48] Robust Vision-Language Models via Tensor Decomposition: A Defense Against Adversarial Attacks (Zhang Qian, 2025) View paper
  - Robustness in Resource-Constrained and Multi-Agent Systems (1 papers)
  - [47] Security and Robustness Challenges of Small Language Models in Autonomous Agent Networks (Micheal, 2025) View paper
- General Adversarial Robustness and Defense Techniques
  - Input Filtering and Dual-Stage Defense Architectures (1 papers)
  - [36] Dual-filtering (DF) schemes for learning systems to prevent adversarial attacks (Dipankar Dasgupta, 2022) View paper
  - Adversarial Purification and Reconstruction Methods (3 papers)
  - [35] Defending Against Model Inversion Attack by Adversarial Examples (Jing Wen, 2021) View paper
  - [38] Defense Against Adversarial Attacks with Efficient Frequency-Adaptive Compression and Reconstruction (Zhonghan Niu, 2023) View paper
  - [46] DiffFilter: Defending Against Adversarial Perturbations With Diffusion Filter (Yong Chen, 2024) View paper
  - Certified Robustness and Causal Intervention Defenses (1 papers)
  - [45] Certified Robustness Against Natural Language Attacks by Causal Intervention (Zhao, 2022) View paper
  - Comprehensive Security Surveys and Threat Landscapes (4 papers)
  - [27] Attack and defense techniques in large language models: A survey and new perspectives (Liao, 2025) View paper
  - [30] Adversarial machine learning for robust cybersecurity: strengthening deep neural architectures against evasion, poisoning, and model-inference attacks (Kalejaiye, 2024) View paper
  - [37] Recent advances in attack and defense approaches of large language models (Cui Jing, 2024) View paper
  - [42] Adversarial attacks on large language models (Jing Zou, 2024) View paper
- Federated and Distributed Learning Security (1 papers)
  - [14] MESAS: Poisoning Defense for Federated Learning Resilient against Adaptive Attackers (Torsten KrauÃ, 2023) View paper

Narrative

Core task: Evaluating robustness of language model defenses against adaptive attacks. The field has organized itself around several major branches that reflect distinct threat models and mitigation strategies. Jailbreak Attack Methods and Characterization explores how adversaries craft prompts to bypass safety guardrails, while Jailbreak Defense Mechanisms develops countermeasures ranging from input filtering to prompt hardening techniques like SmoothLLM[2] and Defensive Prompt Patch[7]. Adaptive Attack Design and Defense Evaluation focuses specifically on the arms race between defenses and attackers who adapt their strategies after observing protective measures, as exemplified by benchmarks such as Jailbreakbench[5] and Jailbreakv Benchmark[1]. Parallel branches address Backdoor Attacks and Defenses, Model Extraction threats, and Domain-Specific Security concerns in areas like code generation or multimodal systems. General Adversarial Robustness techniques and Federated Learning Security round out the taxonomy, highlighting that language model vulnerabilities span multiple deployment contexts and attack surfaces.

Within this landscape, a particularly active line of work examines whether defenses remain effective when attackers can observe and circumvent them—a challenge central to Adaptive Attack Design and Defense Evaluation. Attacker Moves Second[0] sits squarely in this branch, emphasizing the need to test defenses under adaptive threat models where adversaries iteratively refine attacks. This contrasts with static evaluation frameworks and aligns closely with Prompt Injection Evaluation[44], which similarly stresses realistic adversarial conditions. Compared to works like Robust Prompt Optimization[6] or SelfDefend[8] that propose specific defense mechanisms, Attacker Moves Second[0] focuses on the evaluation methodology itself, arguing that many defenses fail when subjected to adaptive scrutiny. The broader tension across these branches revolves around whether defenses can achieve robustness guarantees or merely raise the bar for attackers, with ongoing questions about generalization across attack types and the computational cost of maintaining security under evolving threats.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. A Critical Evaluation of Defenses against Prompt Injection Attacks

Authors: Jia Yuqi, Yuqi Jia, Liu Yu-pei, Zedian Shao, Jia, et al. (14 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Large Language Models (LLMs) are vulnerable to prompt injection attacks, and several defenses have recently been proposed, often claiming to mitigate these attacks successfully. However, we argue that existing studies lack a principled approach to evaluating these defenses. In this paper, we argue the need to assess defenses across two critical dimensions: (1) effectiveness, measured against both existing and adaptive prompt injection attacks involving diverse target and injected prompts, and (2...

Relationship Analysis

Both papers belong to the same taxonomy category of designing adaptive attack frameworks to rigorously test LLM defense robustness. They share substantial overlap in their core mission: demonstrating that existing defenses against jailbreaks and prompt injections fail when evaluated against stronger, adaptive attacks rather than static benchmarks. The key difference is that the original paper presents a unified adaptive attack framework (gradient-based, RL, search-based, and human red-teaming) and breaks 12 defenses, while the candidate paper focuses on establishing principled evaluation methodology for defenses, emphasizing the need to assess both effectiveness against adaptive attacks and general-purpose utility preservation, with detailed case studies on fewer defenses (StruQ, SecAlign, Instruction Hierarchy, PromptGuard, Attention Tracker).

Contributions Analysis

Overall novelty summary. The paper proposes an adaptive attack framework to rigorously evaluate language model defenses against jailbreaks and prompt injections. It resides in the 'Adaptive Attack Frameworks Against LLM Defenses' leaf, which contains only two papers total, indicating a relatively sparse research direction within the broader taxonomy. This positioning suggests the work addresses a recognized but underexplored gap: systematically testing whether defenses withstand adversaries who tailor their strategies after observing protective measures, rather than relying on static attack benchmarks.

The taxonomy reveals that most defense research concentrates in sibling branches such as 'Jailbreak Defense Mechanisms' (with subtopics covering input smoothing, prompt-based strategies, and adversarial training) and 'General Adversarial Robustness' (focusing on filtering and purification methods). The paper's leaf sits under 'Adaptive Attack Design and Defense Evaluation,' which explicitly excludes static benchmarks and general defense surveys. Neighboring leaves address character-level perturbations and baseline defense evaluations, but the adaptive attack framework approach directly targets the evaluation methodology gap rather than proposing new defenses or attack primitives.

Among 29 candidates examined across three contributions, none were found to clearly refute the paper's claims. The first contribution (adaptive attack framework) examined 10 candidates with zero refutable overlaps; the second (comprehensive evaluation of 12 defenses) also examined 10 candidates with no refutations; the third (evaluation recommendations) examined 9 candidates, again with no refutations. This suggests that within the limited search scope—primarily top-K semantic matches and citation expansion—the specific combination of adaptive optimization techniques (gradient descent, reinforcement learning, random search, human-guided exploration) applied systematically to bypass diverse defenses has not been extensively documented in prior work.

Based on the limited literature search covering 29 candidates, the work appears to occupy a distinct position emphasizing evaluation rigor over defense innovation. The taxonomy structure shows that while defense mechanisms and static attack benchmarks are well-populated, the adaptive evaluation methodology remains comparatively sparse. However, the search scope does not guarantee exhaustive coverage of all relevant adversarial evaluation studies, particularly those in adjacent security domains or unpublished concurrent work.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Adaptive attack framework for evaluating LLM defenses

Description: The authors propose a unified adaptive attack framework that systematically applies and scales general optimization techniques (gradient descent, reinforcement learning, random search, and human-guided exploration) to evaluate LLM defenses. This framework is designed to counter specific defense mechanisms rather than using fixed or weak attacks.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Black-box Optimization of LLM Outputs by Asking for Directions

URL: View paper

Brief Assessment

Black Box Optimization[64] focuses on optimizing LLM outputs by exploiting their ability to express comparative confidence through binary comparisons, not on creating a unified adaptive attack framework that systematically applies multiple optimization techniques (gradient descent, RL, random search, human-guided exploration) to evaluate defenses.

2. Certifying LLM Safety against Adversarial Prompting

URL: View paper

Brief Assessment

Certifying LLM Safety[61] focuses on certified defense mechanisms with provable safety guarantees against adversarial prompts, not on developing adaptive attack frameworks for evaluating defenses.

3. Robust Prompt Optimization for Defending Language Models Against Jailbreaking Attacks

URL: View paper

Brief Assessment

Robust Prompt Optimization[6] focuses on defending against jailbreaking attacks through optimization-based defensive suffixes, not on creating a general adaptive attack framework for evaluating defenses. The paper proposes RPO as a defense mechanism that incorporates adversarial objectives during optimization, rather than presenting a unified framework for systematically applying multiple optimization techniques to evaluate defenses.

4. Advprompter: Fast adaptive adversarial prompting for llms

URL: View paper

Brief Assessment

Advprompter[57] focuses on generating adversarial prompts for jailbreaking attacks using a trained LLM model, not on providing a general framework for evaluating defenses using multiple optimization techniques (gradient descent, RL, random search, human-guided exploration) as described in the original contribution.

5. Defending against alignment-breaking attacks via robustly aligned llm

URL: View paper

Brief Assessment

Robustly Aligned LLM[51] focuses on defending against alignment-breaking attacks via a robust alignment checking function, not on developing a unified adaptive attack framework for evaluating defenses. The paper evaluates its defense against existing attacks (GCG, AutoDAN, TAP) rather than proposing a systematic framework for adaptive attack evaluation.

6. AGD: Adversarial Game Defense Against Jailbreak Attacks in Large Language Models

URL: View paper

Brief Assessment

AGD[63] is a defense method that uses adversarial game theory to balance helpfulness and harmlessness in LLMs, not an adaptive attack framework for evaluating defenses. The candidate focuses on defending against jailbreak attacks through internal representation steering and adversarial training, which is fundamentally different from the original paper's contribution of systematically applying optimization techniques to evaluate defense robustness.

7. Gandalf the red: Adaptive security for llms

URL: View paper

Brief Assessment

Gandalf Red[62] focuses on a dynamic security-utility threat model (D-SEC) for LLM applications with crowd-sourced red-teaming, not on a unified adaptive attack framework using gradient descent, RL, random search, and human-guided exploration as optimization techniques to systematically evaluate defenses.

8. Adversarial Training for Large Neural Language Models

URL: View paper

Brief Assessment

Adversarial Training Neural[60] focuses on adversarial training methods to improve model robustness during pre-training and fine-tuning, not on developing adaptive attack frameworks for evaluating defenses. The paper proposes ALUM for training robust models, whereas the original contribution is about systematically evaluating existing defenses using adaptive attacks.

9. Efficient adversarial training in llms with continuous attacks

URL: View paper

Brief Assessment

Continuous Attacks Training[59] focuses on continuous adversarial training in embedding space for robustness, not on developing adaptive attack frameworks for evaluating defenses using multiple optimization techniques (gradient descent, RL, random search, human-guided exploration).

10. Jailbreaking Leading Safety-Aligned LLMs with Simple Adaptive Attacks

URL: View paper

Brief Assessment

[Final Audit Failure] The model insisted on a refutation claim but failed to provide verifiable evidence after multiple retries. Marked as cannot_refute for safety. Please manually verify the candidate text.

Contribution 2: Comprehensive evaluation exposing weaknesses in 12 recent defenses

Description: The authors systematically evaluate 12 recently proposed defenses against jailbreaks and prompt injections using their adaptive attacks. They demonstrate that most defenses can be bypassed with over 90% success rate, contradicting the near-zero attack success rates reported in the original defense papers.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Systematic Testing of Security-Related Vulnerabilities in LLM-Based Applications

URL: View paper

Brief Assessment

Systematic Security Testing[69] focuses on general security vulnerability testing methodologies for LLM applications, not specifically on adaptive attacks against jailbreak/prompt injection defenses or comprehensive evaluation of defense mechanisms.

2. Design Patterns for Securing LLM Agents against Prompt Injections

URL: View paper

Brief Assessment

Design Patterns Security[72] focuses on proposing principled design patterns for building secure AI agents, not on evaluating existing defenses. The candidate does not demonstrate prior comprehensive evaluation work that would refute the original paper's novelty claim.

3. Evolving security in llms: A study of jailbreak attacks and defenses

URL: View paper

Brief Assessment

Evolving Security Study[70] evaluates only 3 defenses (Goal Prioritization, LlamaGuard, Smooth-LLM) against 4 attack methods, not 12 defenses. The scope and scale differ significantly from the original paper's systematic evaluation of 12 defenses with adaptive attacks achieving >90% ASR.

4. PromptBench: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts

URL: View paper

Brief Assessment

PromptBench[65] focuses on evaluating LLM robustness to adversarial prompts (typos, synonyms, perturbations in instructions), not on evaluating defenses against jailbreaks and prompt injections. The candidate paper does not address defense mechanisms or their weaknesses.

5. "Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models

URL: View paper

Brief Assessment

Do Anything Now[67] focuses on characterizing in-the-wild jailbreak prompts and evaluating their effectiveness against LLMs, not on systematically evaluating defense mechanisms. The candidate does not demonstrate prior work on comprehensive defense evaluation.

6. Certifying LLM Safety against Adversarial Prompting

URL: View paper

Brief Assessment

Certifying LLM Safety[61] proposes a defense method (erase-and-check) with certifiable guarantees rather than evaluating existing defenses. It does not systematically expose weaknesses in multiple defense mechanisms.

7. A Comprehensive Study of Jailbreak Attack versus Defense for Large Language Models

URL: View paper

Brief Assessment

Comprehensive Jailbreak Study[18] evaluates 9 attack techniques and 7 defense techniques on 3 models (Vicuna, Llama, GPT-3.5), while the original paper evaluates 12 defenses using adaptive attacks. The candidate focuses on comparing existing techniques rather than demonstrating adaptive attacks that bypass defenses with >90% success rates as claimed in the original work.

8. Evaluating prompt injection safety in large language models using the promptbench dataset

URL: View paper

Brief Assessment

Prompt Injection Safety[66] focuses on evaluating two specific models (Anthropic Claude and Mistral Large) using the PromptBench dataset for prompt injection safety, rather than systematically evaluating multiple defenses with adaptive attacks as described in the original contribution.

9. Evaluating prompt extraction vulnerabilities in commercial large language models

URL: View paper

Brief Assessment

Prompt Extraction Vulnerabilities[68] focuses on prompt extraction attacks against commercial LLMs, not on evaluating jailbreak and prompt injection defenses. The candidate addresses a different security concern (extracting system prompts) rather than bypassing safety defenses.

10. Prompt injection attack against llm-integrated applications

URL: View paper

Brief Assessment

Prompt Injection Attack[71] focuses on attacking LLM-integrated applications through prompt injection techniques, not on evaluating defenses against jailbreaks and prompt injections. The candidate paper does not systematically evaluate multiple defenses or demonstrate their weaknesses.

Contribution 3: Lessons and recommendations for robust defense evaluation

Description: The authors provide four key lessons for the community: static evaluations are misleading, automated attacks are effective but insufficient, human red-teaming remains valuable, and model-based auto-raters can be unreliable. They argue that defense evaluations must incorporate adaptive attackers with substantial computational resources to be convincing.

This contribution was assessed against **9 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. SmoothLLM: Defending Large Language Models Against Jailbreaking Attacks

URL: View paper

Brief Assessment

SmoothLLM[2] focuses on a specific defense mechanism (randomized smoothing) against jailbreaking attacks, not on establishing general evaluation principles for defenses. The paper does not provide systematic recommendations for how the community should evaluate defenses against adaptive attacks.

2. Proactive defense against LLM Jailbreak

URL: View paper

Brief Assessment

Proactive Defense[55] focuses on a novel proactive defense mechanism that generates spurious responses to mislead attackers, rather than providing systematic lessons about defense evaluation methodology. The candidate does not address evaluation standards, adaptive attack requirements, or the role of human red-teaming in assessing defenses.

3. Self-Evaluation as a Defense Against Adversarial Attacks on LLMs

URL: View paper

Brief Assessment

Self Evaluation Defense[53] focuses on a specific defense mechanism (self-evaluation) rather than providing general lessons about defense evaluation methodology. The candidate does not discuss adaptive attackers, computational resources for evaluation, or provide recommendations for the broader community on evaluation practices.

4. A Critical Evaluation of Defenses against Prompt Injection Attacks

URL: View paper

Brief Assessment

[Final Audit Failure] The model insisted on a refutation claim but failed to provide verifiable evidence after multiple retries. Marked as cannot_refute for safety. Please manually verify the candidate text.

5. Defending against alignment-breaking attacks via robustly aligned llm

URL: View paper

Brief Assessment

Robustly Aligned LLM[51] proposes a specific defense mechanism (RA-LLM) and evaluates it against various attacks. While it discusses adaptive attacks in Section 6, it does not provide systematic lessons or recommendations for the community on how to conduct robust defense evaluations with adaptive attackers.

6. Baseline Defenses for Adversarial Attacks Against Aligned Language Models

URL: View paper

Brief Assessment

[Final Audit Failure] The model insisted on a refutation claim but failed to provide verifiable evidence after multiple retries. Marked as cannot_refute for safety. Please manually verify the candidate text.

7. LLM Self Defense: By Self Examination, LLMs Know They Are Being Tricked

URL: View paper

Brief Assessment

LLM Self Examination[56] focuses on a specific defense mechanism (using LLMs to screen their own responses for harmful content) rather than providing comprehensive recommendations for evaluating defenses against adaptive attacks. The candidate does not address evaluation methodology or adaptive attacker considerations that are central to the original paper's contribution.

8. Adversarial tuning: Defending against jailbreak attacks for llms

URL: View paper

Brief Assessment

Adversarial Tuning[40] focuses on a two-stage adversarial tuning framework for defending LLMs against jailbreak attacks through adversarial prompt generation and fine-tuning. It does not provide systematic lessons or recommendations about defense evaluation methodology against adaptive attackers.

9. AutoDefense: Multi-Agent LLM Defense against Jailbreak Attacks

URL: View paper

Brief Assessment

AutoDefense[52] focuses on a multi-agent defense framework against jailbreak attacks, not on establishing evaluation methodologies or recommendations for defense robustness assessment. The paper does not address adaptive attack evaluation standards or provide lessons about defense evaluation practices.

Appendix: Text Similarity Detection

Textual similarity detection checked 28 papers and found 2 similarity segment(s) across 1 paper(s).

The following **1 paper(s)** were detected to have high textual similarity with the original paper. These may represent different versions of the same work, duplicate submissions, or papers with substantial textual overlap. Readers are advised to verify these relationships independently.

1. Robust Prompt Optimization for Defending Language Models Against Jailbreaking Attacks

Detected in: Contribution: contribution_1

⚠ **Note**: This paper shows substantial textual similarity with the original paper. It may be a different version, a duplicate submission, or contain significant overlapping content. Please review carefully to determine the nature of the relationship.

References

- [0] The Attacker Moves Second: Stronger Adaptive Attacks Bypass Defenses Against LLM Jailbreaks and Prompt Injections View paper
- [1] Jailbreakv: A benchmark for assessing the robustness of multimodal large language models against jailbreak attacks View paper
- [2] SmoothLLM: Defending Large Language Models Against Jailbreaking Attacks View paper
- [3] Defending pre-trained language models as few-shot learners against backdoor attacks View paper
- [4] Robustness of Large Language Models Against Adversarial Attacks View paper
- [5] Jailbreakbench: An open robustness benchmark for jailbreaking large language models View paper
- [6] Robust Prompt Optimization for Defending Language Models Against Jailbreaking Attacks View paper
- [7] Defensive Prompt Patch: A Robust and Generalizable Defense of Large Language Models against Jailbreak Attacks View paper
- [8] {SelfDefend}:{LLMs} can defend themselves against jailbreaking in a practical manner View paper
- [9] Enhancing robustness of LLM-driven multi-agent systems through randomized smoothing View paper
- [10] Latent-space adversarial training with post-aware calibration for defending large language models against jailbreak attacks View paper
- [11] Defensive Prompt Patch: A Robust and Interpretable Defense of LLMs against Jailbreak Attacks View paper

- [12] Prediction Poisoning: Towards Defenses Against DNN Model Stealing Attacks View paper
- [13] Defending large language models against jailbreak attacks via layer-specific editing View paper
- [14] MESAS: Poisoning Defense for Federated Learning Resilient against Adaptive Attackers View paper
- [15] Baseline Defenses for Adversarial Attacks Against Aligned Language Models View paper
- [16] Query Provenance Analysis: Efficient and Robust Defense Against Query-Based Black-Box Attacks View paper
- [17] Jailbreak Attacks and Defenses Against Large Language Models: A Survey View paper
- [18] A Comprehensive Study of Jailbreak Attack versus Defense for Large Language Models View paper
- [19] Beyond surface-level patterns: An essence-driven defense framework against jailbreak attacks in llms View paper
- [20] Large Language Models for Code: Security Hardening and Adversarial Testing View paper
- [21] TED-LaST: Towards Robust Backdoor Defense Against Adaptive Attacks View paper
- [22] ModelShield: Adaptive and Robust Watermark Against Model Extraction Attack View paper
- [23] Exploring Clean Label Backdoor Attacks and Defense in Language Models View paper
- [24] Causality based front-door defense against backdoor attack on language models View paper
- [25] From LLMs to MLLMs to Agents: A Survey of Emerging Paradigms in Jailbreak Attacks and Defenses within LLM Ecosystem View paper
- [26] The Tower of Babel Revisited: Multilingual Jailbreak Prompts on Closed-Source Large Language Models View paper
- [27] Attack and defense techniques in large language models: A survey and new perspectives View paper
- [28] Attacks and Defenses for Large Language Models on Coding Tasks View paper
- [29] Visual Adversarial Examples Jailbreak Large Language Models View paper
- [30] Adversarial machine learning for robust cybersecurity: strengthening deep neural architectures against evasion, poisoning, and model-inference attacks View paper
- [31] A Comprehensive Defense Framework Against Model Extraction Attacks View paper
- [32] Shieldlearner: A new paradigm for jailbreak attack defense in llms View paper
- [33] Do Methods to Jailbreak and Defend LLMs Generalize Across Languages? View paper
- [34] SpeechGuard: Exploring the adversarial robustness of multimodal large language models View paper
- [35] Defending Against Model Inversion Attack by Adversarial Examples View paper
- [36] Dual-filtering (DF) schemes for learning systems to prevent adversarial attacks View paper
- [37] Recent advances in attack and defense approaches of large language models View paper
- [38] Defense Against Adversarial Attacks with Efficient Frequency-Adaptive Compression and Reconstruction View paper
- [39] Certifying Language Model Robustness with Fuzzed Randomized Smoothing: An Efficient Defense Against Backdoor Attacks View paper
- [40] Adversarial tuning: Defending against jailbreak attacks for llms View paper
- [41] Differential Robustness in Transformer Language Models: Empirical Evaluation Under Adversarial Text Attacks View paper
- [42] Adversarial attacks on large language models View paper
- [43] Rapid Response: Mitigating LLM Jailbreaks with a Few Examples View paper
- [44] A Critical Evaluation of Defenses against Prompt Injection Attacks View paper
- [45] Certified Robustness Against Natural Language Attacks by Causal Intervention View paper
- [46] DifFilter: Defending Against Adversarial Perturbations With Diffusion Filter View paper
- [47] Security and Robustness Challenges of Small Language Models in Autonomous Agent Networks View paper
- [48] Robust Vision-Language Models via Tensor Decomposition: A Defense Against Adversarial Attacks View paper
- [49] Towards Backdoor Attacks and Defense in Robust Machine Learning Models View paper
- [50] TrojanZoo: Towards Unified, Holistic, and Practical Evaluation of Neural Backdoors View paper
- [51] Defending against alignment-breaking attacks via robustly aligned llm View paper
- [52] AutoDefense: Multi-Agent LLM Defense against Jailbreak Attacks View paper
- [53] Self-Evaluation as a Defense Against Adversarial Attacks on LLMs View paper
- [54] Adaptive Attacks Break Defenses Against Indirect Prompt Injection Attacks on LLM Agents View paper
- [55] Proactive defense against LLM Jailbreak View paper
- [56] LLM Self Defense: By Self Examination, LLMs Know They Are Being Tricked View paper
- [57] Advprompter: Fast adaptive adversarial prompting for llms View paper
- [58] Jailbreaking Leading Safety-Aligned LLMs with Simple Adaptive Attacks View paper
- [59] Efficient adversarial training in llms with continuous attacks View paper
- [60] Adversarial Training for Large Neural Language Models View paper
- [61] Certifying LLM Safety against Adversarial Prompting View paper
- [62] Gandalf the red: Adaptive security for llms View paper
- [63] AGD: Adversarial Game Defense Against Jailbreak Attacks in Large Language Models View paper
- [64] Black-box Optimization of LLM Outputs by Asking for Directions View paper
- [65] PromptBench: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts View paper
- [66] Evaluating prompt injection safety in large language models using the promptbench dataset View paper
- [67] "Do Anything Now": Characterizing and Evaluating In-The-Wild Jailbreak Prompts on Large Language Models View paper
- [68] Evaluating prompt extraction vulnerabilities in commercial large language models View paper
- [69] Systematic Testing of Security-Related Vulnerabilities in LLM-Based Applications View paper
- [70] Evolving security in llms: A study of jailbreak attacks and defenses View paper
- [71] Prompt injection attack against llm-integrated applications View paper
- [72] Design Patterns for Securing LLM Agents against Prompt Injections View paper